Data Science

- Turning **data** into something meaningful
- **Science** of uncertainty
- Quintessential **interdisciplinary** science

Data Science Skillset

- Statistics, mathematics and IT skills (e.g. programming)

Data Science Skillset

- **Statistics**, mathematics and IT skills (e.g. **programming**)

Data Science Skillset

- **Statistics**, mathematics and IT skills (e.g. **programming**)
- Logical thinker
- Problem solver
- Good communicator

What does the term,

“statistics”,

mean to you ?

What is / are Statistics?

What is / are Statistics?

A statistic:

Science of statistics:

What is / are Statistics?

A statistic[*]: any quantity computed from sample data

Science of statistics:
collecting, classifying, summarizing, organizing, analyzing, estimation and interpretation of information

[*] Terminology also used for function to calculate the summary quantity

Role of Statistics

Field of statistics deals with the collection, presentation, analysis, and use of data to:

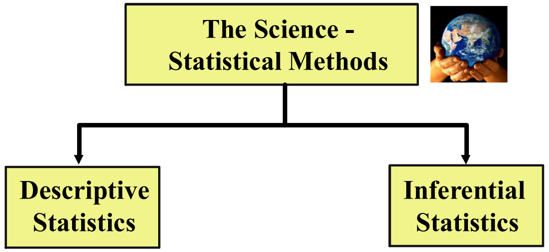
- make decisions
- solve problems
- design products and processes

It is the science of uncertainty

Role of Probability

•Probability provides the **framework** for the study and application of statistics

Statistical Methods



https://www.nuigalway.ie/adult-learning/about-us/didyouknow/

https://visual.ly/community/infographic/animals/shark-attack

Descriptive Statistics: *Science of summarizing data, numerically and graphically…*

Analysis methods applicable depends on the variable being measured and the research questions which you are trying to answer …

Inferential Statistics: *science of using the **information in your sample** to say (i.e. to "**infer**") something **about the population** of interest*

Thinking Challenge

Suppose the student newspaper is interested in what proportion of NUI Galway students pay rent

and

the average amount of rent paid

How would you find out?

Breakdown the question...

What is the individual / experimental unit?

What is the population of interest?

What are the variables of interest?

What types are these variables?

What are the parameters of interest?

How would you collect the data?

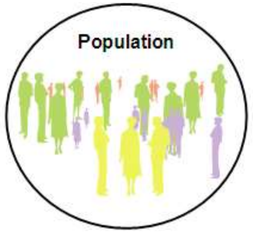
What are the observations for the variables?

How would you summarise these observations?

Some important terms:

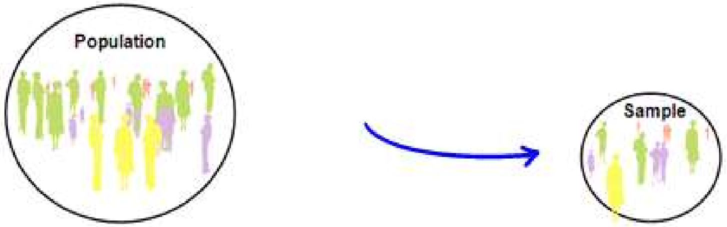
An ***experimental unit / individual*** is a single object upon which we collect data,
e.g. person,
thing,
transaction,
event.

Some important terms:



A ***population***
is a collection of
experimental units/individuals
that we are interested in studying.
e.g. people,
things,
transactions,
events

Some important terms:



Some important terms:

A **sample**
is a subset of experimental units /
individuals from the population.
e.g. people,
things,
transactions,
events

Some important terms:

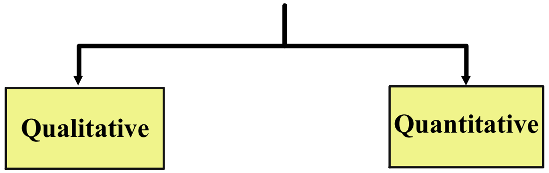
A **variable** is a characteristic or property of an individual experimental unit.

examples:

height
grade score
account balance
gender (m/f/non-binary),
letter grade (A, B, C, etc.),
Likert scale (agree, neutral, disagree, etc.)

Types of variable:

A **variable** is a characteristic or property of an individual experimental unit



May be measured, or more generally "observed", on each individual

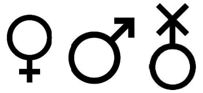
Qualitative Data:

Classified into categories, can be **ordered**:

- Grade achieved in ST2001

or **unordered**:

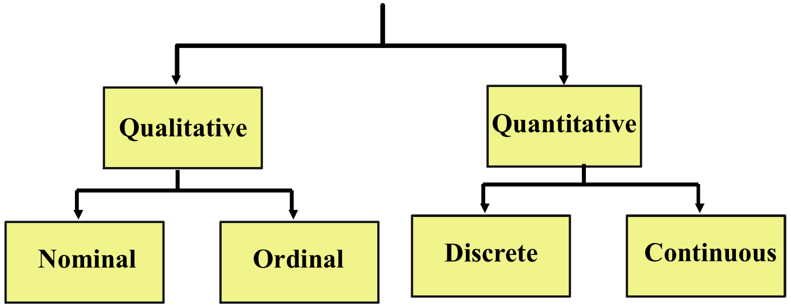
- Gender of each employee at a company



- Method of payment (cash, cheque, credit card)

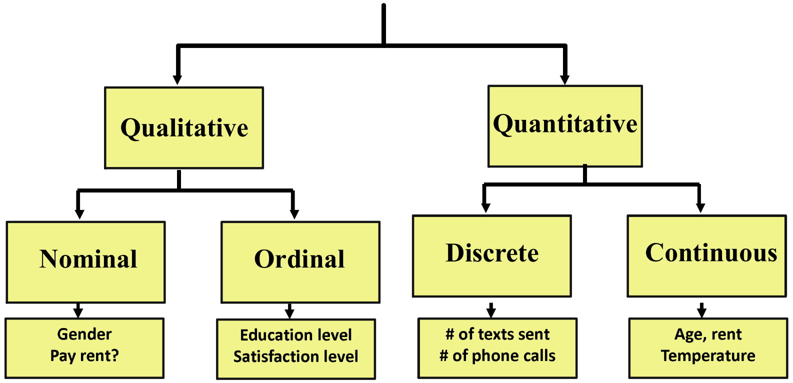
Types of variable:

A **variable** is a characteristic or property of an individual experimental unit.



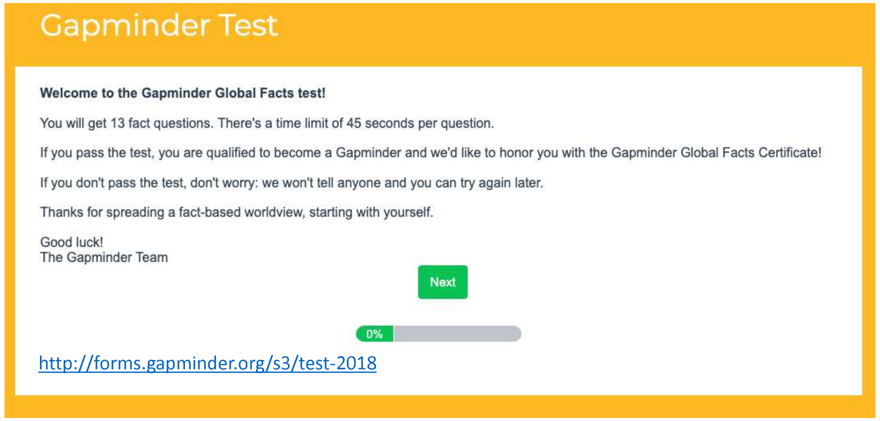
Types of variable:

A **variable** is a characteristic or property of an individual experimental unit.



Gapminder Data: https://www.gapminder.org/

The Gapminder Foundation is a Swedish NGO which promotes sustainable global development by increased use and understanding of statistics about social, economic and environmental development



Gapminder Test

Welcome to the Gapminder Global Facts test!

You will get 13 fact questions. There's a time limit of 45 seconds per question.

If you pass the test, you are qualified to become a Gapminder and we'd like to honor you with the Gapminder Global Facts Certificate!

If you don't pass the test, don't worry: we won't tell anyone and you can try again later.

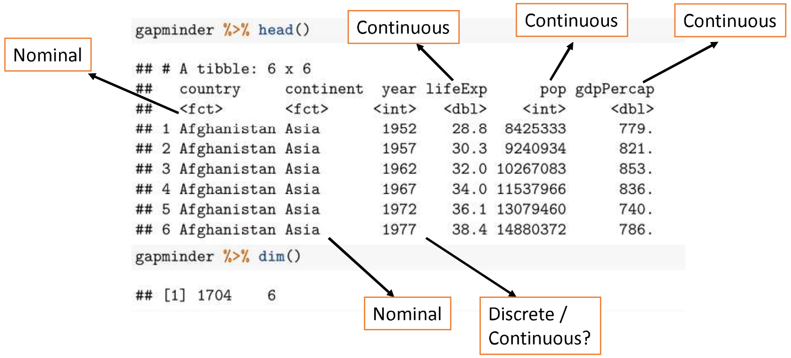
Thanks for spreading a fact-based worldview, starting with yourself.

Good luck!
The Gapminder Team

http://forms.gapminder.org/s3/test-2018

- What is the *typical observation*?

Gapminder Data



```
gapminder %>% head()

## # A tibble: 6 x 6
##   country     continent  year lifeExp      pop gdpPercap
##   <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
## 1 Afghanistan Asia       1952    28.8  8425333      779.
## 2 Afghanistan Asia       1957    30.3  9240934      821.
## 3 Afghanistan Asia       1962    32.0 10267083      853.
## 4 Afghanistan Asia       1967    34.0 11537966      836.
## 5 Afghanistan Asia       1972    36.1 13079460      740.
## 6 Afghanistan Asia       1977    38.4 14880372      786.
gapminder %>% dim()

## [1] 1704    6
```

Nominal

Continuous

Continuous

Continuous

Nominal

Discrete / Continuous?

- What is the *typical observation*?
- Is there much *variation/spread* between individuals in the dataset?

- What is the *typical observation*?
- Is there much *variation/spread* between individuals in the dataset?
- How are the observations distributed over all individuals in the group – i.e. what is the shape or *distribution*?

- What is the *typical observation*?
- Is there much *variation/spread* between individuals in the dataset?
- How are the observations distributed over all individuals in the group – i.e. what is the shape or *distribution*?

- Are there any values lying outside of the range where the majority of the dataset values lie – *outliers*?

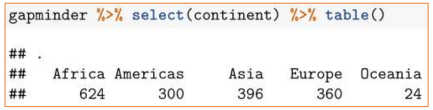
- What is the *typical observation*?
- Is there much *variation/spread* between individuals in the dataset?
- How are the observations distributed over all individuals in the group – i.e. what is the shape or *distribution*?
- Are there any values lying outside of the range where the majority of the dataset values lie – *outliers*?

Summarising data (variables) can be done **numerically**, with appropriate summaries, or **graphically**, with appropriate plots

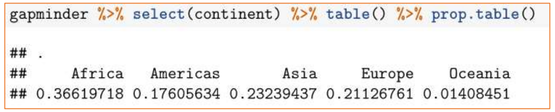
Summarising Categorical Data

- **Numerical Summary:** frequency count and percentage

Continent	Frequency	Proportion
Africa	624	0.36619718
Americas	300	0.17605634
Asia	396	0.23239437
Europe	360	0.21126761
Oceania	24	0.01408451



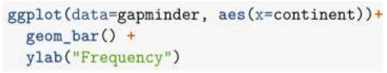
```
gapminder %>% select(continent) %>% table()

## .
##   Africa Americas     Asia   Europe  Oceania
##      624      300      396      360       24
```



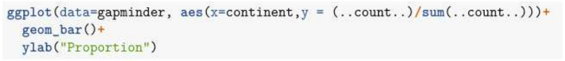
```
gapminder %>% select(continent) %>% table() %>% prop.table()

## .
##     Africa   Americas       Asia     Europe    Oceania
## 0.36619718 0.17605634 0.23239437 0.21126761 0.01408451
```

Summarising Categorical Data

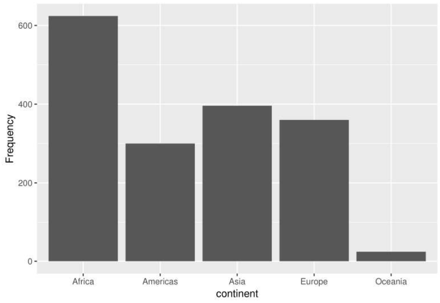
- Graphical summary: bar chart, pie chart

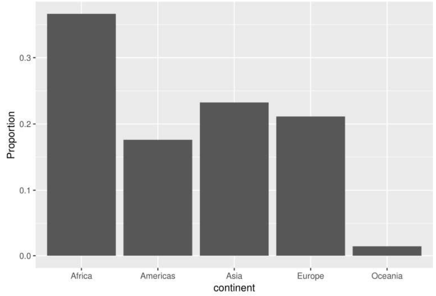


```
ggplot(data=gapminder, aes(x=continent))+
  geom_bar() +
  ylab("Frequency")
```



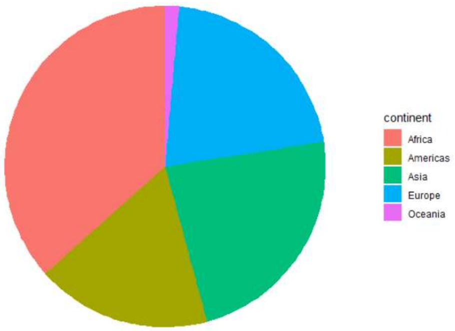
```
ggplot(data=gapminder, aes(x=continent,y = (..count..)/sum(..count..)))+
  geom_bar()+
  ylab("Proportion")
```





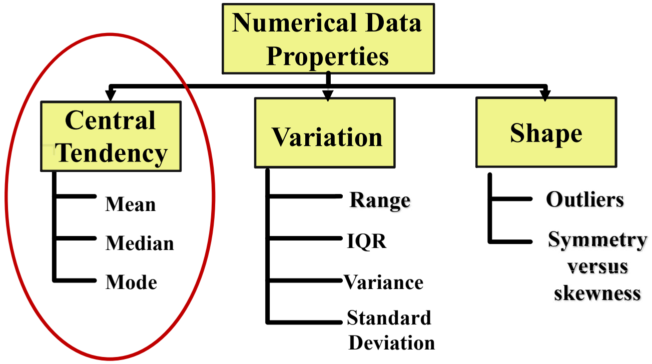
Summarising Categorical Data

- Graphical summary: bar chart, pie chart



Advice: don't use pie charts
People find determining angles very difficult
Easier to understand lengths/heights

Summarising Continuous Data



Numerical Data Properties

Central Tendency
- Mean
- Median
- Mode

Variation
- Range
- IQR
- Variance
- Standard Deviation

Shape
- Outliers
- Symmetry versus skewness

Numerical summary of typical value:

Definition

Suppose that the observations in a sample are $x_1, x_2, \ldots, x_n$. The **sample mean**, denoted by $\bar{x}$, is

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Sensitive to extreme values

Given that the observations in a sample are $x_1, x_2, \ldots, x_n$, arranged in **increasing order** of magnitude, the sample median is

$$\tilde{x} = \begin{cases} x_{(n+1)/2}, & \text{if } n \text{ is odd}, \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}), & \text{if } n \text{ is even}. \end{cases}$$

NOT Sensitive to extreme values

Mode is the most frequent observation in a dataset.

Example

Data: breaking strength of wire in kilograms

220 214 222 218 223 210 223 210 227 225 212

- Find the median:
  - Order the data from lowest to highest

210 210 212 214 218 220 222 223 223 225 227

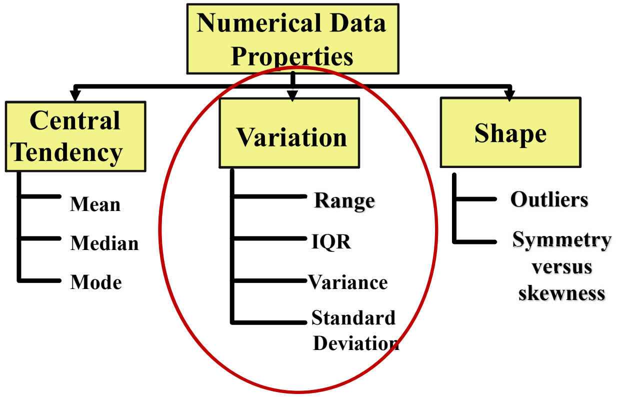
Median

- Find the Mean:

$$Mean = \frac{220 + 214 + \ldots + 222}{11} = 218.5455$$

- Mode is 210 and 223, as both have been repeated twice

Summarising Continuous Data



Numerical Data Properties

Central Tendency
- Mean
- Median
- Mode

Variation
- Range
- IQR
- Variance
- Standard Deviation

Shape
- Outliers
- Symmetry versus skewness

Numerical Summary of Spread

- Range = ***maximum - minimum***

Examples:

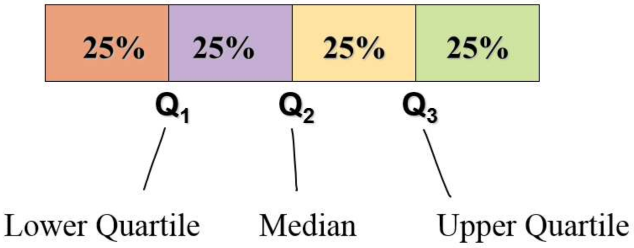
- 1, 2, 5, 8, 10 gives range of 10 – 1 = 9
- 1, 5, 5, 5, 10 also gives range of 9

- Clearly the range is poor measure of spread
- Also badly affected by outliers

Numerical Summary of Spread

- Interquartile range (IQR = Q_3 - Q_1)
- Middle 50% range of data, so is robust to outliers

Split ordered data into 4 quarters



25% 25% 25% 25%

Q_1 Q_2 Q_3

Lower Quartile Median Upper Quartile

Tukey’s Method for IQR (lots of others)

Data: breaking strength of wire in kilograms
220 214 222 218 223 210 223 210 227 225 212

Put data in ascending order:

210 210 212 214 218 220 222 223 223 225 227

Q_1 = 213 Median Q_3 = 223

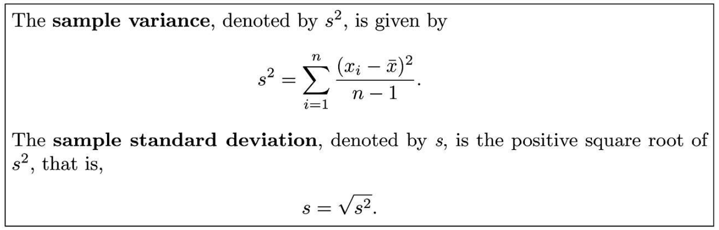
Lower (Upper) quartile is median of lower (upper) 50% of data including the median

$$\text{IQR} = Q_3 - Q_1 = 223 – 213 = 10$$

Numerical Summary of Spread

- Common measure of spread is the standard deviation, which takes into account how far *each* data value is from the mean
- A deviation is the distance of a datapoint from the mean
- Since the sum of all the deviations would be zero, we square each deviation and find an average (of sorts) of them (called the **variance**)
- We the square-root this average squared deviation… **Why?**

Definition



The **sample variance**, denoted by s^2, is given by

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}.$$

The **sample standard deviation**, denoted by s, is the positive square root of s^2, that is,

$$s = \sqrt{s^2}.$$

Sample Standard Deviation

- In same units as original variable
  - So preferable to sample variance, which is in squared units
- But… it is sensitive to outliers

Example

Data: breaking strength of wire in kilograms

220 214 222 218 223 210 223 210 227 225 212

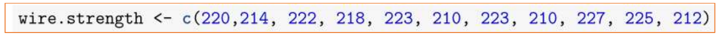
- **Find the sample variance**
- **Find the sample standard deviation**

$$\bar{x} = 218.5455$$

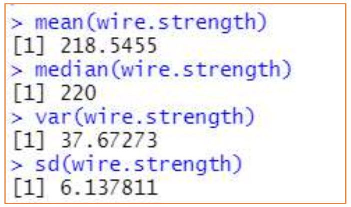
$$Sample\ Variance = s^2 = \frac{(220-218.5455)^2+(214-218.5455)^2+\cdots+(222-218.5455)^2}{11-1} = 37.67273$$

$$Sample\ Standard\ deviation = s = \sqrt{Sample\ Variance} = \sqrt{37.67273} = 6.1378$$

Numerical Summary in R: Vector

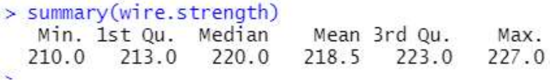


```
wire.strength <- c(220,214, 222, 218, 223, 210, 223, 210, 227, 225, 212)
```



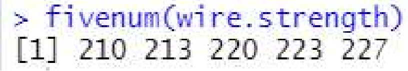
```
> mean(wire.strength)
[1] 218.5455
> median(wire.strength)
[1] 220
> var(wire.strength)
[1] 37.67273
> sd(wire.strength)
[1] 6.137811
```

summary() function uses a different formula for quartiles



```
> summary(wire.strength)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  210.0   213.0   220.0   218.5   223.0   227.0 
```

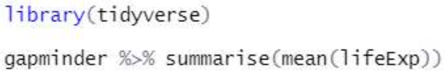
fivenum() function uses Tukey's method for Q_1 and Q_3, called the five number summary



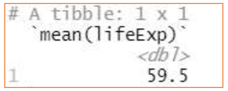
```
> fivenum(wire.strength)
[1] 210 213 220 223 227
```

Numerical Summary in R:

Calculate the **mean** of life expectancy for gapminder data:

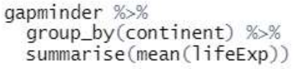


```
library(tidyverse)

gapminder %>% summarise(mean(lifeExp))
```

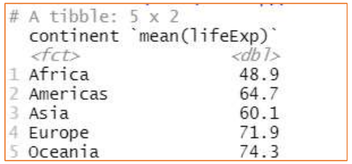


```
# A tibble: 1 x 1
  `mean(lifeExp)`
            <dbl>
1            59.5
```

Calculate the **mean** of life expectancy for different continents:

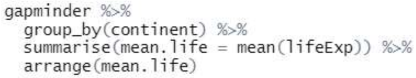


```
gapminder %>%
  group_by(continent) %>%
  summarise(mean(lifeExp))
```

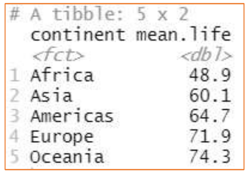


```
# A tibble: 5 x 2
  continent `mean(lifeExp)`
  <fct>               <dbl>
1 Africa               48.9
2 Americas             64.7
3 Asia                 60.1
4 Europe               71.9
5 Oceania              74.3
```

arrange

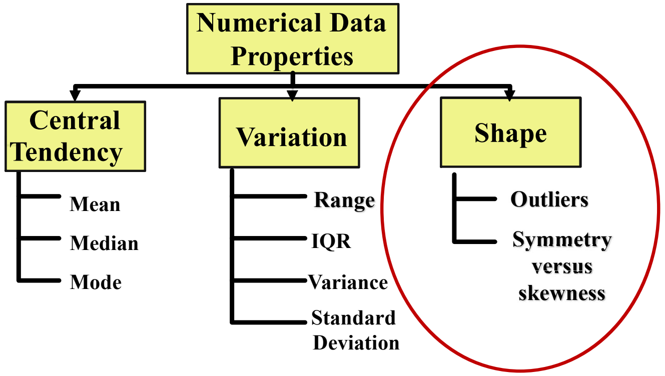


```
gapminder %>%
  group_by(continent) %>%
  summarise(mean.life = mean(lifeExp)) %>%
  arrange(mean.life)
```



```
# A tibble: 5 x 2
  continent mean.life
  <fct>         <dbl>
1 Africa         48.9
2 Asia           60.1
3 Americas       64.7
4 Europe         71.9
5 Oceania        74.3
```

Summarising Continuous Data



Numerical Data Properties

Central Tendency
- Mean
- Median
- Mode

Variation
- Range
- IQR
- Variance
- Standard Deviation

Shape
- Outliers
- Symmetry versus skewness

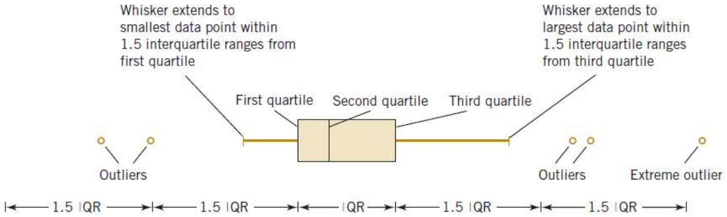
Summarising Continuous Data: Shape

- Graphical summary: boxplot, histogram

Boxplot

- A boxplot is a graphical display showing center, spread, shape, and outliers.
- It displays the 5-number summary:

  min, Q_1, *median*, Q_3, and *max*

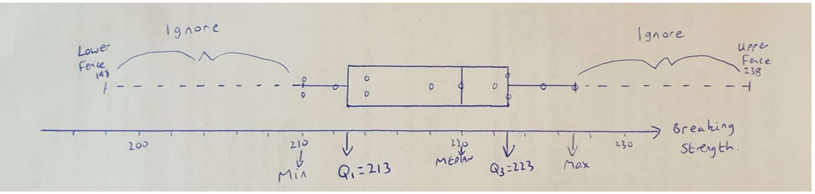

Whisker extends to smallest data point within 1.5 interquartile ranges from first quartile
Whisker extends to largest data point within 1.5 interquartile ranges from third quartile
First quartile
Second quartile
Third quartile
Outliers
Outliers
Extreme outlier
1.5 IQR
1.5 IQR
IQR
1.5 IQR
1.5 IQR

Boxplot of Breaking Length

Data: breaking strength of wire in kilograms
220 214 222 218 223 210 223 210 227 225 212

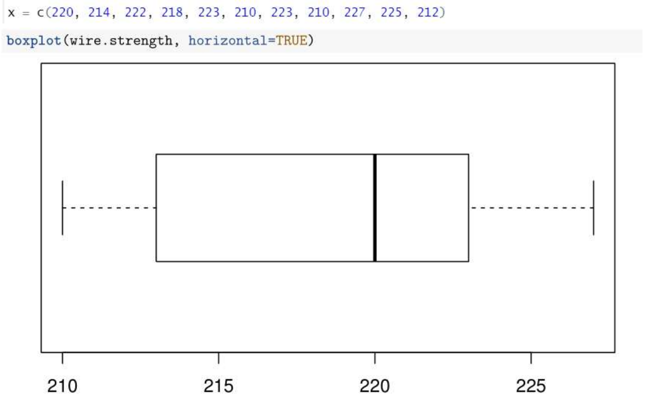
```
Variable           Minimum      Q1   Median      Q3   Maximum
Breaking Length     210.00  213.00   220.00  223.00    227.00
```

Upper fence: Q_3 + 1.5 IQR = 223 + 1.5 × 10 = 238
Lower fence: Q_1 – 1.5 IQR = 213 – 1.5 × 10 = 198



Think about a garden “fence” and closest ball is within your gardenn!

Graphical Summary in R: `boxplot()`



```
x = c(220, 214, 222, 218, 223, 210, 223, 210, 227, 225, 212)
```

```
boxplot(wire.strength, horizontal=TRUE)
```

- Note: `boxplot()` function in `R` gives exactly same result
- Other functions / software may use different method to calculate the quartiles (and/or fences)
- Usually these differences are minor so can be ignored

Histograms

✓ Useful to show general shape, location and spread of data values – representation by ***area***

Construction

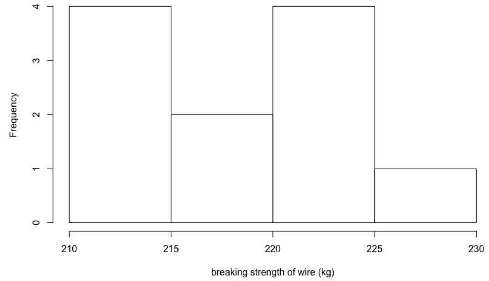
- Determine range of data – *minimum*, *maximum*
- Split into convenient intervals (or bins)
- Usually use 5 to 15 intervals
- Count number of observations in each interval - *frequency*

Histogram of Breaking Length

Data: breaking strength of wire in kilograms

220 214 222 218 223 210 223 210 227 225 212

- **Find the minimum and maximum**
- **Make classes of width 5 starting from minimum**
- **Count the frequency**
- **Plot the histogram!**



Shape of the data

When talking about the shape of the data, make sure to address the following three questions:

1. Does the histogram have a single, central hump or several well separated bumps?
2. Is the histogram or boxplot symmetric? Or more spread out in one direction, i.e. skewed
3. Any unusual features? e.g. outliers, spikes

Features to look for

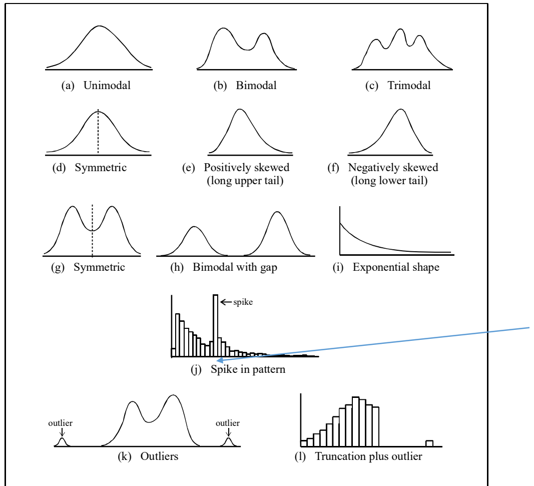


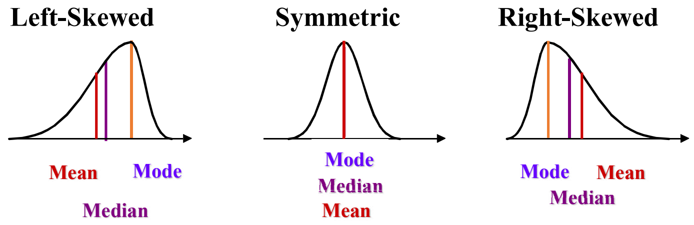
Figure 2.3.10 Features to look for in histograms and stem-and-leaf plots.

From *Chance Encounters* by C.J. Wild and G.A.F. Seber. © John Wiley & Sons, 2000.

e.g. minimum value for free postage!!!

Remember the mean, median and mode ?

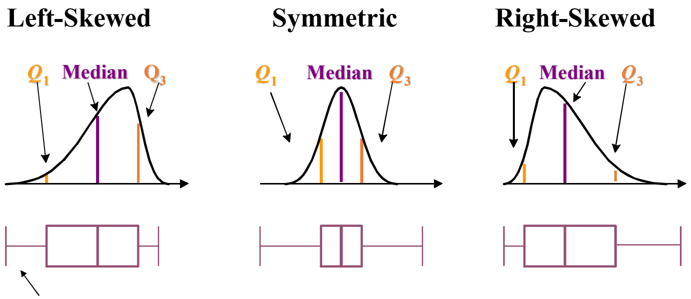
The mean is the average data value,



The value of the mean is strongly affected by skewness and outliers, - more so than the median.

Shape & Box Plot

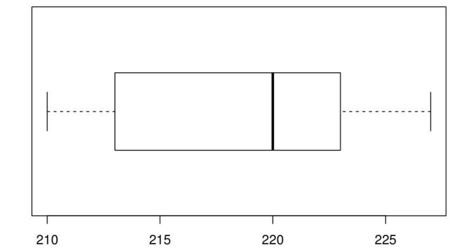
These shapes can also be seen in the boxplots



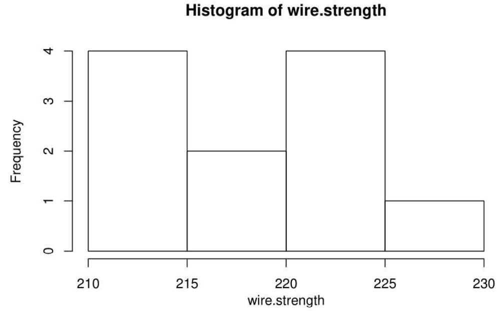
Left skewed - Longer tail on left than right, median may not be central in the box.

Graphical Summary in R: Vector

```
boxplot(wire.strength, horizontal=TRUE)
```



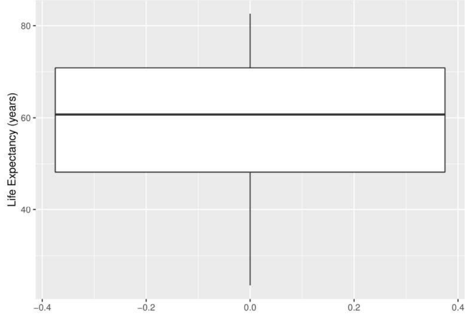
```
hist(wire.strength)
```



Graphical Summary in R: Dataframe

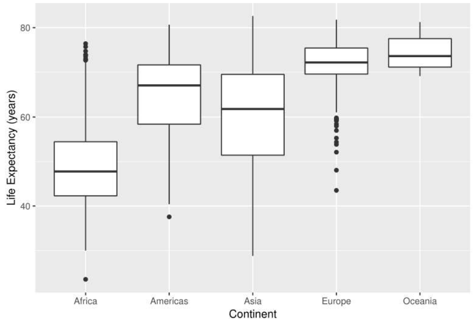
Plot the **boxplot** of life expectancy for gapminder data:

```
ggplot(gapminder, aes(y = lifeExp)) +
  geom_boxplot() +
  ylab("Life Expectancy (years)")
```



Plot the **boxplot** of life expectancy for different continents:

```
ggplot(gapminder, aes(x = continent, y = lifeExp)) +
  geom_boxplot() +
  ylab("Life Expectancy (years)") +
  xlab("Continent")
```



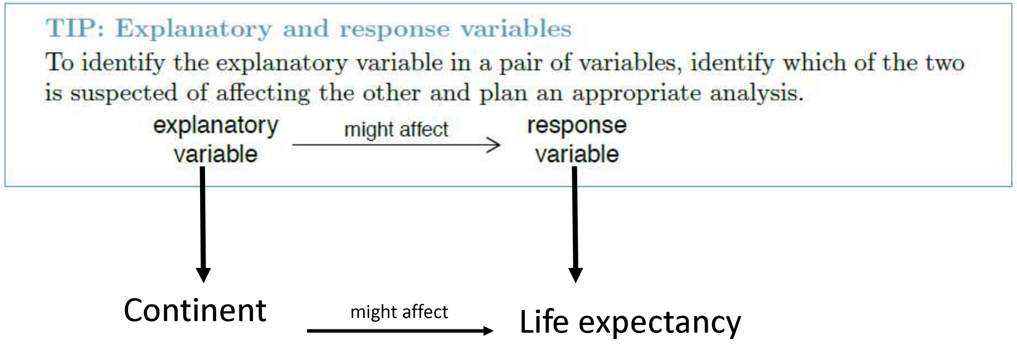
Explanatory and response variables

TIP: Explanatory and response variables

To identify the explanatory variable in a pair of variables, identify which of the two is suspected of affecting the other and plan an appropriate analysis.

explanatory variable → might affect → response variable

Explanatory and response variables



TIP: Explanatory and response variables

To identify the explanatory variable in a pair of variables, identify which of the two is suspected of affecting the other and plan an appropriate analysis.

Graphical summaries of data

- Depends on the variable of interest
- Categorical response variable: barchart (n or %) or pie chart
- Categorical response variable with an explanatory variable: grouped barchart
- Continuous response variable: histogram, boxplot, density plot
- Continuous response variable with an explanatory variable: grouped boxplot

Using R

- R statistical computing and visualisation software
- Free open source package,
- Commonly used software for statistics

- 18,000+ contributed packages / libraries
- Lots of tutorials online
- Lots of sources of online help

DataCamp

Search

THE SMARTEST WAY TO

Learn Data Science Online

The skills people and businesses need to succeed are changing. No matter where you are in your career or what field you work in, you will need to understand the language of data. With DataCamp, you learn data science today and apply it tomorrow.

DataCamp

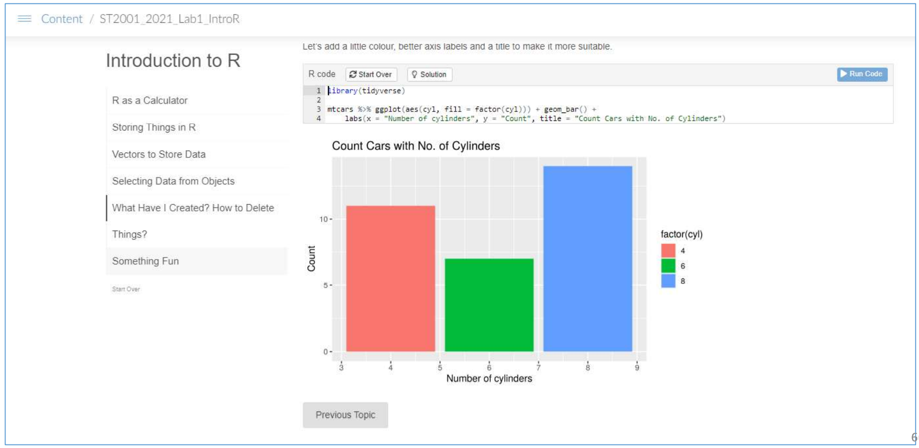
Search

COURSES

- Introduction to Python
- Introduction to R
- Introduction to SQL
- Data Science for Everyone
- Introduction to Data Engineering
- Introduction to Deep Learning in Python

See all courses (338)

A Gentle Start in R



Content / ST2001_2021_Lab1_IntroR

Introduction to R

- R as a Calculator
- Storing Things in R
- Vectors to Store Data
- Selecting Data from Objects
- What Have I Created? How to Delete Things?
- Something Fun

Start Over

Let's add a little colour, better axis labels and a title to make it more suitable.

R code

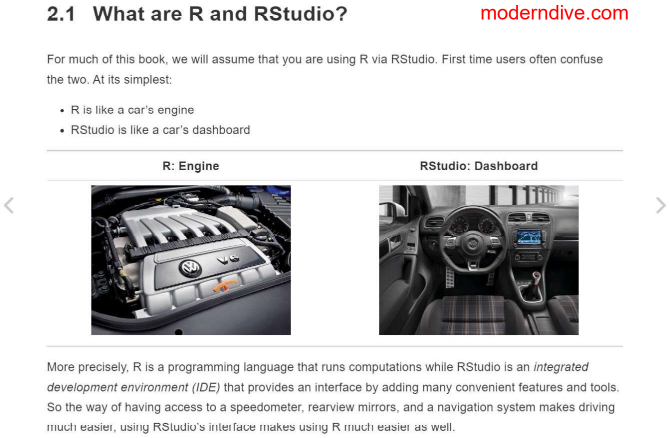
```
library(tidyverse)

mtcars %>% ggplot(aes(cyl, fill = factor(cyl))) + geom_bar() +
    labs(x = "Number of cylinders", y = "Count", title = "Count Cars with No. of Cylinders")
```

cran.r-project.org

+

www.rstudio.com/download

RStudio®



2.1 What are R and RStudio?

moderndive.com

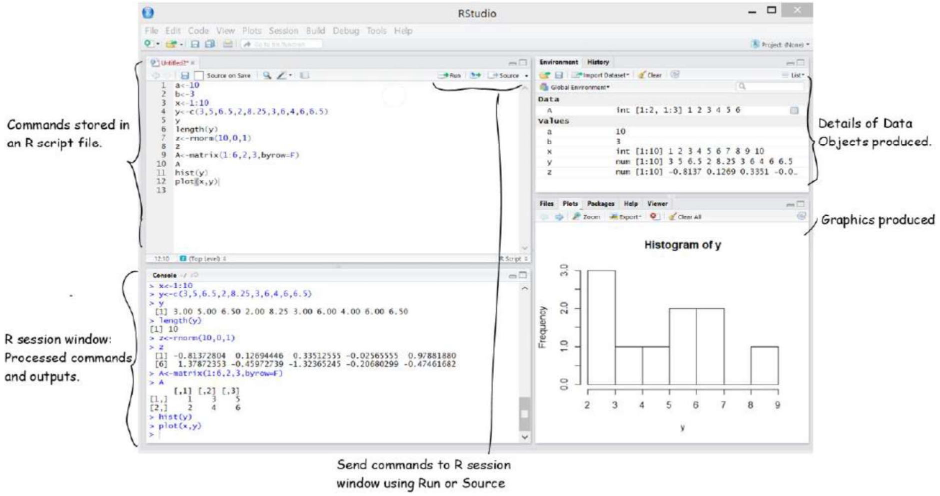
For much of this book, we will assume that you are using R via RStudio. First time users often confuse the two. At its simplest:

- R is like a car's engine
- RStudio is like a car's dashboard

R: Engine	RStudio: Dashboard

More precisely, R is a programming language that runs computations while RStudio is an *integrated development environment (IDE)* that provides an interface by adding many convenient features and tools. So the way of having access to a speedometer, rearview mirrors, and a navigation system makes driving much easier, using RStudio's interface makes using R much easier as well.

RStudio is an integrated development environment for R.


RStudio
File Edit Code View Plots Session Build Debug Tools Help
Untitled2*
Source on Save
Run
Source
Environment History
Import Dataset
Clear
List
Global Environment
Project: (None)
Files Plots Packages Help Viewer
Zoom
Export
Clear All
12:10
(Top level)
R Script
Console ~/
Histogram of y
Frequency
0.0
1.0
2.0
3.0
2
3
4
5
6
7
8
9
y
Commands stored in an R script file.
Details of Data Objects produced.
Graphics produced
R session window: Processed commands and outputs.
Send commands to R session window using Run or Source

```
a<-10
b<-3
x<-1:10
y<-c(3,5,6.5,2,8.25,3,6,4,6,6.5)
y
length(y)
z<-rnorm(10,0,1)
z
A<-matrix(1:6,2,3,byrow=F)
A
hist(y)
plot(x,y)
```

```
Data
  A            int [1:2, 1:3] 1 2 3 4 5 6
Values
  a            10
  b            3
  x            int [1:10] 1 2 3 4 5 6 7 8 9 10
  y            num [1:10] 3 5 6.5 2 8.25 3 6 4 6 6.5
  z            num [1:10] -0.8137 0.1269 0.3351 -0.0…
```

```
> x<-1:10
> y<-c(3,5,6.5,2,8.25,3,6,4,6,6.5)
> y
 [1] 3.00 5.00 6.50 2.00 8.25 3.00 6.00 4.00 6.00 6.50
> length(y)
[1] 10
> z<-rnorm(10,0,1)
> z
 [1] -0.81372804  0.12694446  0.33512555 -0.02565555  0.97881880
 [6]  1.37872353 -0.45972739 -1.32365245 -0.20680299 -0.47461682
> A<-matrix(1:6,2,3,byrow=F)
> A
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
> hist(y)
> plot(x,y)
>
```

Installing R and RStudio

Tutorial in installing R and RStudio on your computer (and key packages):

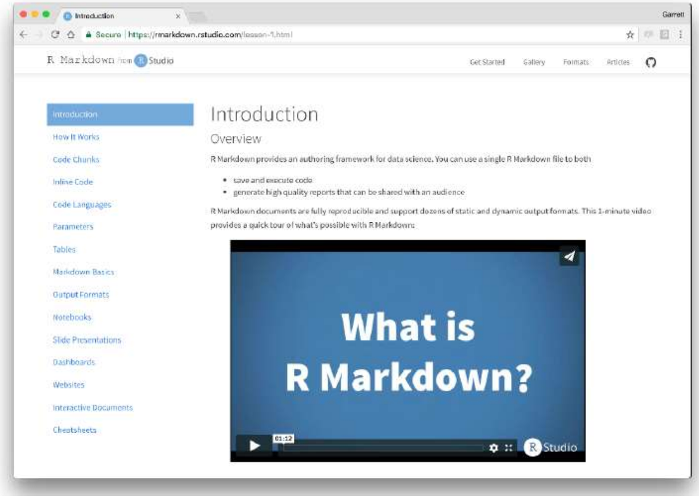
https://jjallaire.shinyapps.io/learnr-tutorial-00-setup/

More instructions videos on Blackboard, but do also google!

Introducing R Markdown

- R Markdown is a file format for making dynamic documents with R
- Written in markdown (an easy-to-write plain text format) and contains:
  - chunks of embedded R code (data management, summaries, graphics, tables, analysis and interpretation)
  - all in the **one** document
- Document can be **knit**ted to html, pdf, word and many other formats!

https://rmarkdown.rstudio.com/lesson-1.html



Introduction

Overview

R Markdown provides an authoring framework for data science. You can use a single R Markdown file to both

- save and execute code
- generate high quality reports that can be shared with an audience

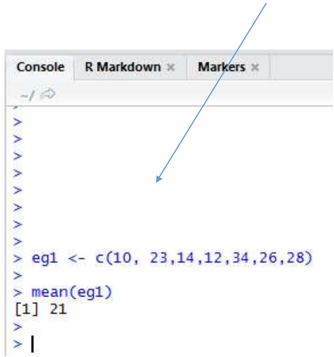
R Markdown documents are fully reproducible and support dozens of static and dynamic output formats. This 1-minute video provides a quick tour of what's possible with R Markdown:

Key Benefits of R Markdown

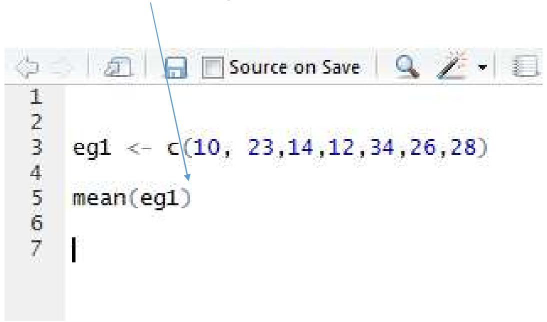
- R Markdown makes it easy to produce statistical reports with code, analysis, outputs and write-up all in one place
- Perfect for reproducible research!
- Easy to convert to different document types

https://github.com/rstudio/cheatsheets/raw/master/rmarkdown.pdf

Drawback of terminal and R script?

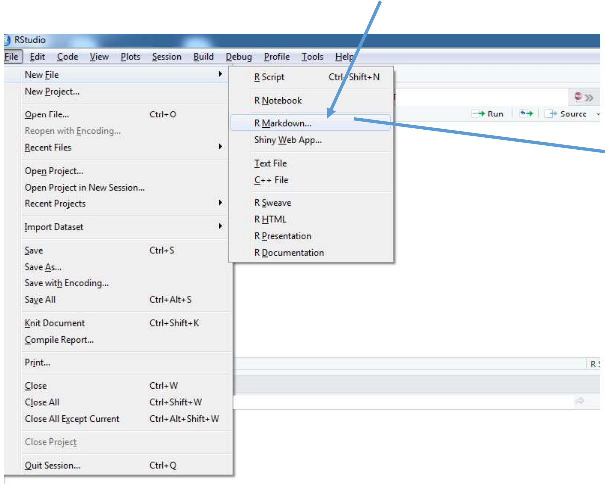


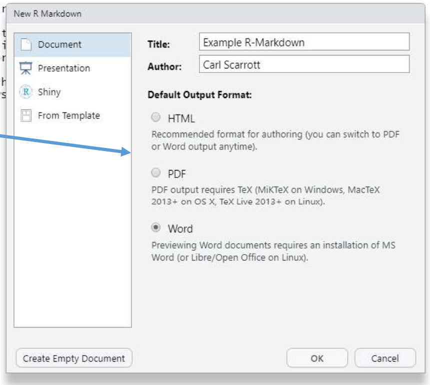
```
>
>
>
>
>
>
>
>
>
> eg1 <- c(10, 23,14,12,34,26,28)
>
> mean(eg1)
[1] 21
>
> |
```



```

eg1 <- c(10, 23,14,12,34,26,28)

mean(eg1)

|
```

Creating R Markdown Document





Title: Example R-Markdown

Author: Carl Scarrott

Default Output Format:

○ HTML

Recommended format for authoring (you can switch to PDF or Word output anytime).

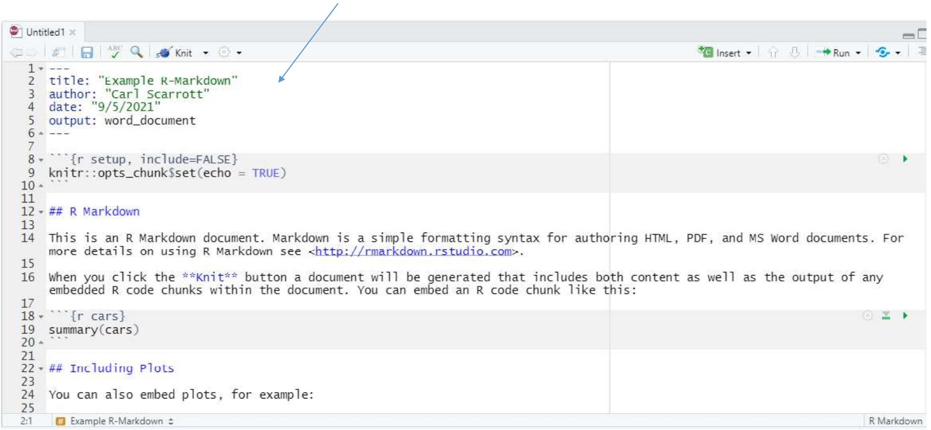
○ PDF

PDF output requires TeX (MiKTeX on Windows, MacTeX 2013+ on OS X, TeX Live 2013+ on Linux).

● Word

Previewing Word documents requires an installation of MS Word (or Libre/Open Office on Linux).

Basic R Markdown Document



~~~
---
title: "Example R-Markdown"
author: "Carl Scarrott"
date: "9/5/2021"
output: word_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
~~~

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

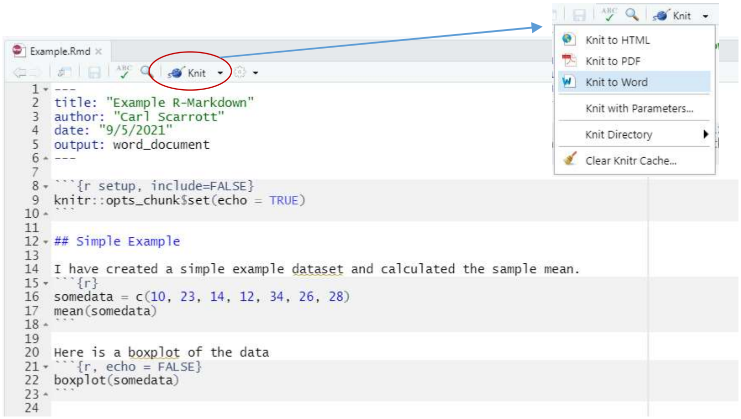
```{r cars}
summary(cars)
```

Including Plots

You can also embed plots, for example:

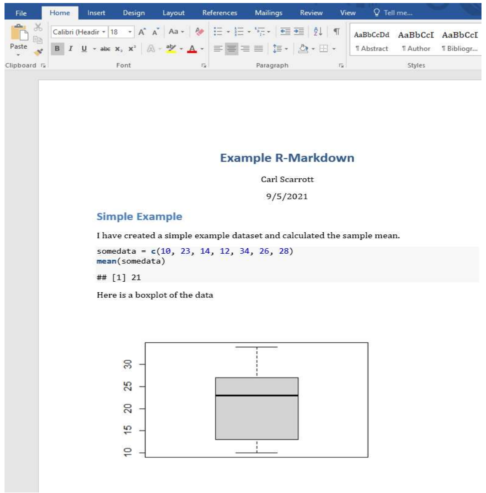
```
```

Edit and “knit” Document



~~~
---
title: "Example R-Markdown"
author: "Carl Scarrott"
date: "9/5/2021"
output: word_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

## Simple Example

I have created a simple example dataset and calculated the sample mean.
```{r}
somedata = c(10, 23, 14, 12, 34, 26, 28)
mean(somedata)
```

Here is a boxplot of the data
```{r, echo = FALSE}
boxplot(somedata)
```

~~~

R Markdown knitted to Word



Example R-Markdown

Carl Scarrott

9/5/2021

Simple Example

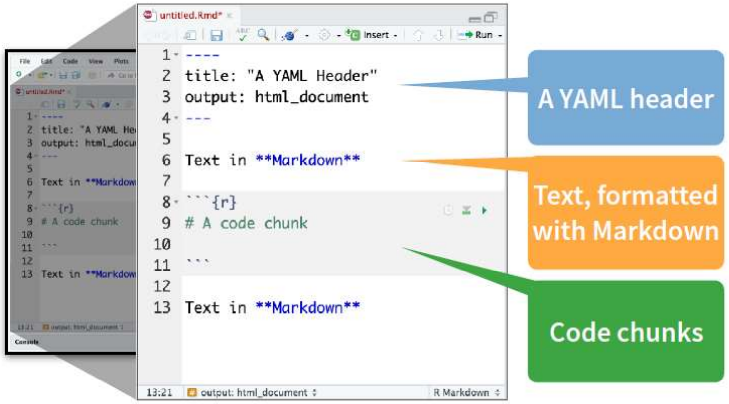
I have created a simple example dataset and calculated the sample mean.

```
somedata = c(10, 23, 14, 12, 34, 26, 28)
mean(somedata)

## [1] 21
```

Here is a boxplot of the data

Structure

R Markdown documents contain **three types of content**



~~~
---
title: "A YAML Header"
output: html_document
---

Text in **Markdown**

```{r}
# A code chunk

```

Text in **Markdown**
~~~

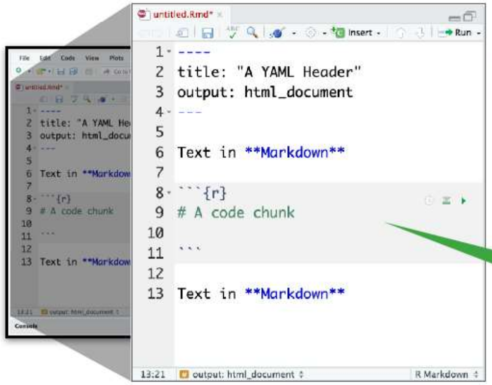
A YAML header

Text, formatted with Markdown

Code chunks

Code Chunks

Write and execute code in a **chunk**. Insert with



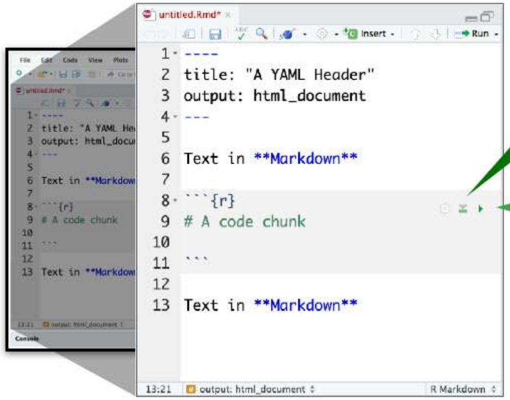
~~~
---
title: "A YAML Header"
output: html_document
---

Text in **Markdown**

```{r}
# A code chunk

```

Text in **Markdown**
~~~

Code chunks

- **Command + Opt + I** ()
- **Control + Alt + I** ()
- GUI Insert button
- Typing out the tick marks

Code Chunks

Write and execute code in a **chunk**.



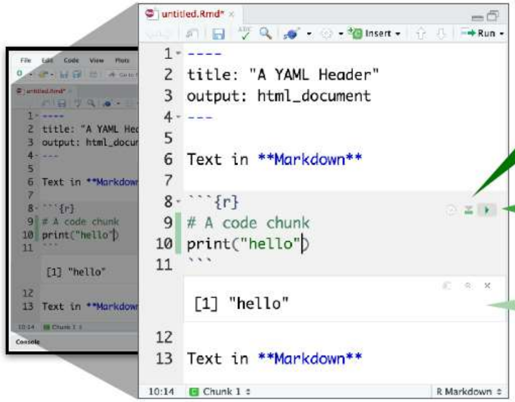
~~~
----
title: "A YAML Header"
output: html_document
---

Text in **Markdown**

```{r}
# A code chunk

```

Text in **Markdown**
~~~

Click to run all code chunks above

Click to run code in chunk

Code Chunks

Write and execute code in a **chunk**.



~~~
----
title: "A YAML Header"
output: html_document
---

Text in **Markdown**

```{r}
# A code chunk
print("hello")
```

 [1] "hello"

Text in **Markdown**
~~~

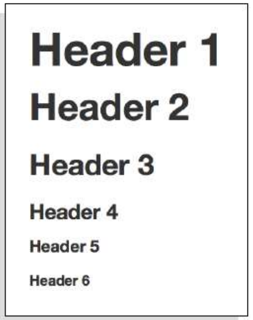
Click to run all code chunks above

Click to run code in chunk

Code result

Headers

```
# Header 1
## Header 2
### Header 3
#### Header 4
##### Header 5
###### Header 6
```



Header 1

Header 2

Header 3

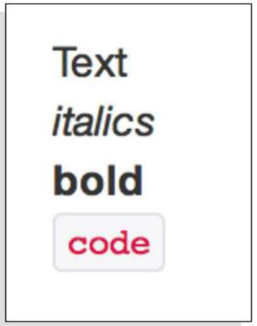
Header 4

Header 5

Header 6

Text

```
Text
_italics_
__bold__
`code`
```



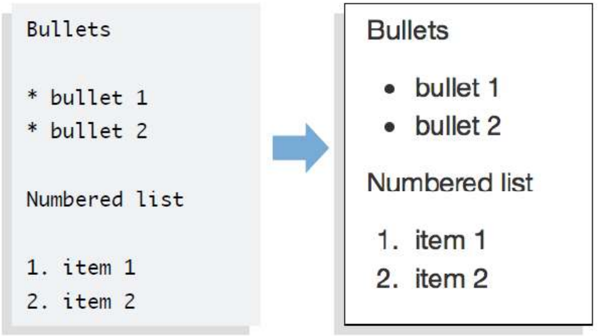
Text

italics

bold

`code`

Lists



```
Bullets

* bullet 1
* bullet 2

Numbered list

1. item 1
2. item 2
```

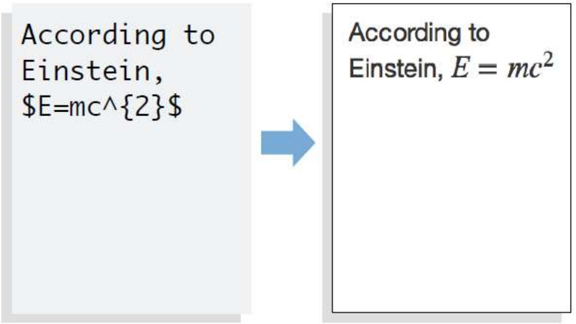
Bullets

- bullet 1
- bullet 2

Numbered list

1. item 1
2. item 2

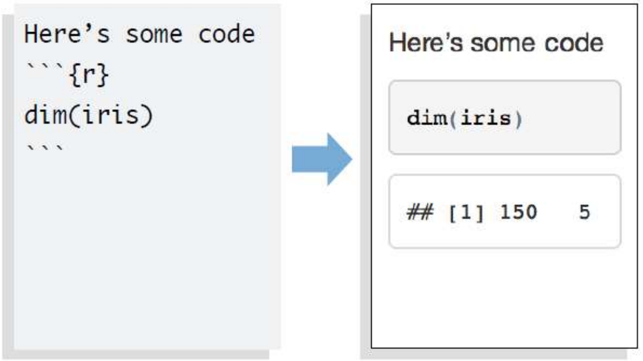
Equations

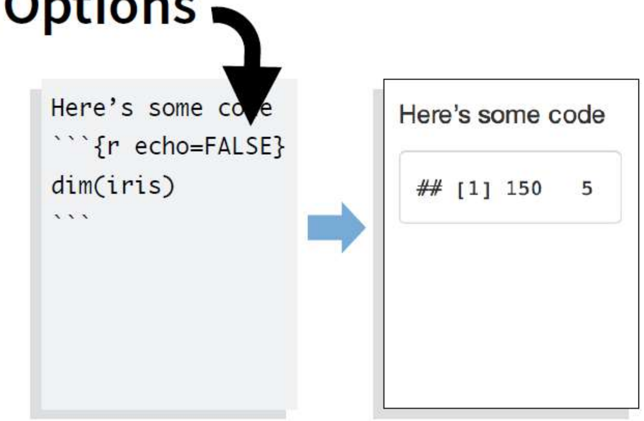


```
According to
Einstein,
$E=mc^{2}$
```

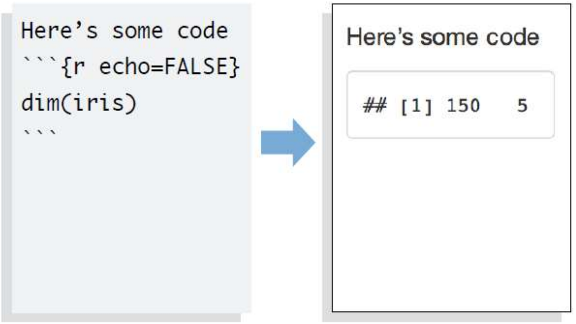
According to Einstein, $E = mc^2$

Code chunks



~~~
Here's some code
```{r}
dim(iris)
~~~
```

Here's some code

```
dim(iris)
```

```
## [1] 150   5
```


Chunk Options



```
Here's some code
```{r echo=FALSE}
dim(iris)
```
```

Here's some code

```
[1] 150   5
```
```

echo = FALSE



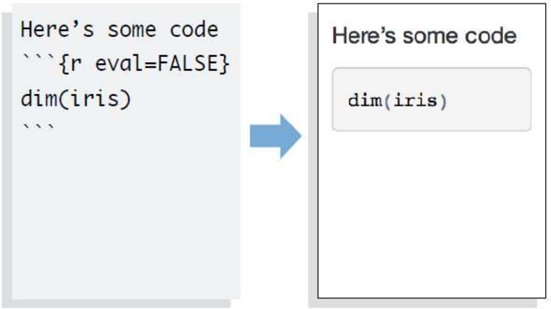
~~~
Here's some code
```{r echo=FALSE}
dim(iris)
```
~~~

Here's some code

```
## [1] 150    5
```

Displays code results, but **not code**

eval = FALSE



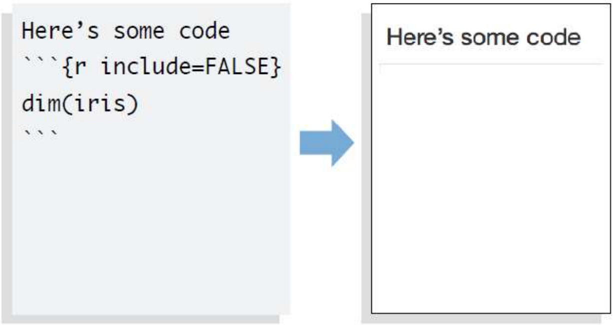
~~~
Here's some code
```{r eval=FALSE}
dim(iris)
```
~~~

Here's some code

```
dim(iris)
```

Displays code, but **not results** (code is not run)

include = FALSE



~~~
Here's some code
```{r include=FALSE}
dim(iris)
```
~~~

Here's some code

Displays **neither code not results** (but code is run)